ETL Project

Question: Where do people smoke the most tobacco? Does weather and population of a State affect the population's smoking habits?

Data Sources:

1. Census Bureau API - population data
2. Web scraping "usa.com/rank" - weather averages for the year
    a. http://www.usa.com/rank/us--average-max-temperature--state-rank.htm?yr=9000&dis=&wist=&plow=&phigh=
    b. http://www.usa.com/rank/us--average-min-temperature--state-rank.htm?yr=9000&dis=&wist=&plow=&phigh=
3. Kaggle csv file - data on smoking behaviour of different States

Transformation: all transformations were done using pandas on jupyter notebook

1. Census API data - main transformation required here was to change the data frame headers to say "State" and "Population"
2. Web scraped weather data - this required a lot of editing, mainly making sure that the data frame was translated properly following "pd.read_html" (i.e. correct headers from html, etc), splitting the column containing "State/Population" data to just "State" (yes, we could have just used population data from this website but we wanted to show off our API skills), turning the temperatures from objects to numeric values, and finally combining the two tables (average max temperature and average min temperature) into one
3. Smoking behaviors csv file - we extracted the data only from 2010, and then performed an inner merge with the census data to make sure extra "States" such as the row named "Nationwide (States, DC, and Territories)" did not appear in our final database. We also removed the % sign from the data and turned them into a numeric values instead of objects
4. Final transformation - after making the table schema on pgAdmin, where the census population table contained the primary key 'state', and the other two tables contained this primary key as a foreign key, we had to make sure that all tables had the same exact States. We then had to change the headers of every dataframe on pandas to match the table headers on pgAdmin.

Final Database:

Our final database (loaded at the end of the file MasterDF.ipynb) contained every State, its population, average minimum temperature and average maximum temperature, and smoking behavior data. We used a SQL database because we wanted a relational database that connected our three tables.